ED458213 2001-04-00 Reliability. ERIC Digest.

ERIC Development Team

www.eric.ed.gov

Table of Contents

If you're viewing this document online, you can click any of the topics below to link directly to that section.

ERIC Identifier: ED458213
Publication Date: 2001-04-00
Author: Rudner, Lawrence M. - Schafer, William D.
Source: ERIC Clearinghouse on Assessment and Evaluation College Park MD.

Reliability. ERIC Digest.

All tests contain error. This is true for tests in the physical sciences and for educational and psychological tests. In measuring length with a ruler, for example, there may be systematic error associated with where the zero point is printed on the ruler and random error associated with your eye's ability to read the marking and extrapolate between the markings. It is also possible that the length of the object can vary over time and environment (e.g., with changes in temperature). One goal in assessment is to keep these errors down to levels that are appropriate for the purposes of the test. High-stakes tests, such as licensure examinations, need to have very little error. Classroom tests can tolerate more error because it is fairly easy to spot and correct mistakes made during the testing process. Reliability focuses only on the degree of errors that are nonsystematic, called random errors.

Reliability has been defined in different ways by different authors. Perhaps the best way to look at reliability is the extent to which the measurements resulting from a test are the result of characteristics of those being measured. For example, reliability has elsewhere been defined as "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker" (Berkowitz, Wolkowitz, Fitch, and Kopriva, 2000). This definition will be satisfied if the scores are indicative of properties of the test takers; otherwise they will vary unsystematically and not be repeatable or dependable.

Reliability can also be viewed as an indicator of the absence of random error when the test is administered. When random error is minimal, scores can be expected to be more consistent from administration to administration.

Technically, the theoretical definition of reliability is the proportion of score variance that is caused by systematic variation in the population of test-takers. This definition is population-specific. If there is greater systematic variation in one population than another, such as in all public school students compared with only eighth-graders, the test will have greater reliability for the more varied population. This is a consequence of how reliability is defined. Reliability is a joint characteristic of a test and examinee group, not just a characteristic of a test. Indeed, reliability of any one test varies from group to group. Therefore, the better research studies will report the reliability for their sample as well as the reliability for norming groups as presented by the test publisher.

This Digest discusses sources of error, several approaches toward estimating reliability, and several ways to increase test reliability.

SOURCES OF ERROR

There are three major sources of error: factors in the test itself, factors in the students taking the test, and scoring factors. Most tests contain a collection of items that represent particular skills. We typically generalize from each item to all items like that item. For example, if a student can solve several problems like 7 times 8, then we may generalize his or her ability to multiply single-digit integers. We also generalize from the collection of items to a broader domain. If a student does well on a test of addition, subtraction, multiplication, and division of fractions, then we may generalize and conclude that the student is able to perform fraction operations. But error may be introduced by the selection of particular items to represent the skills and domains. The particular cross-section of test content that is included in the specific items on the test will vary with each test form, introducing sampling error and limiting the dependability of the test, since we are generalizing to unobserved data, namely, ability across all items that could have been on the test. On basic arithmetic skills, one would expect the content to be fairly similar, making it relatively easy to build a highly reliable test. As the skills and domains become more complex, more errors are likely introduced by sampling of items. Other sources of test error include the effectiveness of the distractors (wrong options) in multiple choice tests, partially correct distractors, multiple correct answers, and difficulty of the items relative to the student's ability.

As human beings, students are not always consistent and also introduce error into the testing process. Whether a test is intended to measure typical or optimal student performance, changes in such things as student's attitudes, health, and sleep may affect the quality of their efforts and thus their test-taking consistency. For example, test takers may make careless errors, misinterpret test instructions, forget test instructions, inadvertently omit test sections, or misread test items.

Scoring errors are a third potential source of error. On objective tests, the scoring is mechanical, and scoring error should be minimal. On constructed-response items, sources of error include clarity of the scoring rubrics, clarity of what is expected of the student, and a host of rater errors. Raters are not always consistent, sometimes change their criteria while scoring, and are subject to biases such as the halo effect, stereotyping, perception differences, leniency/stringency error, and scale shrinkage (see Rudner, 1992).

MEASURES OF RELIABILITY

It is impossible to calculate a reliability coefficient that conforms to the theoretical definition. Recall, the theoretical definition depends on knowing the degree to which a population of examinees vary in their true achievement (or whatever the test measures). But if we knew that, then we wouldn't need the test! Instead, there are several statistics (coefficients) commonly used to estimate the stability of a set of test scores for a group of examinees: test-retest reliability, split-half reliability, measures of internal consistency, and alternate form reliability are the most common.

Test-retest reliability. A test-retest reliability coefficient is obtained by administering the same test twice and correlating the scores. In concept, it is an excellent measure of score consistency because it allows the direct measurement of consistency from administration to administration. This coefficient is not recommended in practice, however, because of its problems and limitations. It requires two administrations of the same test with the same group of individuals. This is expensive and not a good use of people's time. If the time interval is short, people may be overly consistent because they remember some of the questions and their responses. If the interval is long, then the results are confounded with learning and maturation, that is, changes in the persons themselves.

Split-half reliability. As the name suggests, split-half reliability is a coefficient obtained by dividing a test into halves, correlating the scores on each half, and then correcting for length (longer tests tend to be more reliable). The split can be based on odd versus even numbered items, randomly selecting items, or manually balancing content and difficulty. This approach has an advantage in that it only requires a single test administration. Its weakness is that the resultant coefficient will vary as a function of how the test was split. It is also not appropriate on tests in which speed is a factor (that is, where students' scores are influenced by how many items they reached in the allotted time).

Internal consistency. Internal consistency focuses on the degree to which the individual items are correlated with each other and is thus often called homogeneity. Several statistics fall within this category. The best known are Cronbach's alpha, the Kuder-Richardson Formula 20 (KR-20) and the Kuder-Richardson Formula 21 (KR-21). Most testing programs that report data from one administration of a test to students do so using Cronbach's alpha, which is functionally equivalent to KR-20.

The advantages of these statistics are that they only require one test administration and that they do not depend on a particular split of items. The disadvantage is that they are most applicable when the test measures a single skill area.

Requiring only the test mean, standard deviation (or variance), and the number of items, the Kuder-Richardson formula 21 is an extremely simple reliability formula. While it will almost always provide coefficients that are lower than KR-20, its simplicity makes it a very useful estimate of reliability, especially for evaluating some classroom-developed tests. However, it should not be used if the test has items that are scored other than just zero or one.

Where M is the mean, k is the number of items, and v is the test variance.

Alternate-form reliability. Most standardized tests provide equivalent forms that can be used interchangeably. These alternate forms are typically matched in terms of content and difficulty. The correlation of scores on pairs of alternate forms for the same examinees provides another measure of consistency or reliability. Even with the best test and item specifications, each test would contain slightly different content and, as with test-retest reliability, maturation and learning may confound the results. However, the use of different items in the two forms conforms to our goal of including the extent to which item sets contribute to random errors in estimating test reliability.

HOW HIGH SHOULD RELIABILITY BE?

Most large-scale tests report reliability coefficients that exceed .80 and often exceed .90. The questions to ask are 1) What are the consequences of the test? and 2) Is the group used to compute the reported reliability similar to my group?

If the consequences are high, as in tests used for special education placement, high school graduation, and professional certification, then the internal consistency reliability needs to be quite high--at least above .90, preferably above .95. Misclassifications due to measurement error should be kept to a minimum. And please note that no test should ever be used by itself to make an important decision for anyone.

Classroom tests seldom need to have exceptionally high reliability coefficients. As more students master the content, test variability will go down and so will the coefficients from internal measures of reliability. Further, classroom tests don't need exceptionally high reliability coefficients. Teachers see their students all day and have opportunities to gather input from a variety of information sources. Teacher knowledge and judgment, used along with information from the test, provides superior information. If a test is not reliable or it is not accurate for an individual, teachers can and should make the appropriate corrections. A reliability coefficient of .50 or .60 may suffice.

Again, reliability is a joint characteristic of a test and examinee group, not just a characteristic of a test. Thus, reliability also needs to be evaluated in terms of the examinee group. A test with a reliability of .92 when administered to students in 4th, 5th, and 6th grades will not have as high a reliability when administered just to a group of 4th graders. IMPROVING TEST RELIABILITY

Developing better tests with less random measurement error is better than simply documenting the amount of error. Measurement error is reduced by writing items clearly, making the instructions easily understood, adhering to proper test administration, and providing consistent scoring. Because a test is a sample of the desired skills and behaviors, longer tests, which are larger samples, will be more reliable. A one-hour end-of-unit exam will be more reliable than a five-minute pop quiz.

A COMMENT ON SCORING

How should teachers respond when children make careless mistakes on a test? On one hand, teachers want students to learn to follow directions, to think through their work, to check their work, and to be careful. On the other hand, tests are supposed to reflect what a student knows. A low score due to careless mistakes is not the same as a low score due to lack of knowledge.

We believe that especially in the elementary grades, a miserable test due to careless mistakes should not dramatically lower a student's grade for the semester. The semester grade should reflect what the student has achieved, since that is the meaning it will convey to others. We therefore advocate keeping two sets of records, especially in the elementary grades. One set reflects production, and the other reflects achievement. The teacher then has the needed data to apply good judgment in conferencing with parents and for determining semester grades.

REFERENCES AND RECOMMENDED READING

Anastasi, A. (1988). Psychological Testing. New York: MacMillan Publishing Company. Berkowitz, D., Wolkowitz, B., Fitch, R., Kopriva, R. (2000). The Use of Tests as Part of High-Stakes Decision-Making for Students: A Resources Guide for Educators and Policy-Makers. Washington, DC: U.S. Department of Education. [Available online: http://www.ed.gov/offices/OCR/testing/].

Lyman, H. B. (1993). Test Scores and What They Mean. Boston: Allyn and Bacon.

McMillan, J. H. (2001). Essential Assessment Concepts for Teachers and Administrators. Thousand Oaks, CA: Corwin Publishing Company.

Nunnally, J. C. (1967). Psychometric Theory (Chapters 6 and 7). New York: McGraw-Hill Book Company.

Popham, W. James (1998). Classroom Assessment, What Teachers Need to Know. Boston: Allyn and Bacon.

Rudner, Lawrence M. (1992). Reducing errors due to the use of judges. Practical Assessment, Research & Evaluation, 3(3). [Available online: http://ericae.net/pare/getvn.asp?v=3&n=3].

This publication was prepared with funding from the Office of Educational Research and Improvement, U.S. Department of Education, under contract ED99CO0032. The opinions expressed in this report do not necessarily reflect the positions of policies of OERI or the U.S. Department of Education. Permission is granted to copy and distribute this ERIC/AE Digest.

—

Title: Reliability. ERIC Digest.
Document Type: Information Analyses---ERIC Information Analysis Products (IAPs) (071); Information Analyses---ERIC Digests (Selected) in Full Text (073);
Available From: ERIC Clearinghouse on Assessment and Evaluation, 1129 Shriver Laboratory, University of Maryland, College Park, MD 20742. Tel: 800-464-3742 (Toll Free); Web site: http://ericae.net.
Descriptors: Educational Testing, Error of Measurement, Reliability, Scores, Test Results, Testing Problems
Identifiers: ERIC Digests
###

—

▲

[Return to ERIC Digest Search Page]